Supplementary Material for: Dessie: Disentanglement for Articulated 3D Horse Shape and Pose Estimation from Images

Ci Li[1], Yi Yang[1,2], Zehang Weng[1],
Elin Hernlund[3], Silvia Zuffi[4], and Hedvig Kjellström[1,3]

[1] KTH, Sweden {cil, yiya, zehang, hedvig}@kth.se
[2] Scania, Sweden carol-yi.yang@scania.com
[3] SLU, Sweden Elin.Hernlund@slu.se
[4] IMATI-CNR, Italy silvia@mi.imati.cnr.it

This document provides the supplementary materials. Section 1 presents how we employ a diffusion model in DessiePIPE to synthesize realistic UV texture maps. In Section 2, we offer additional qualitative results using out-of-domain images. Section 3 offers visualization of the DINO key features for DinoHMR. Section 4 offers additional analysis regarding training Dessie with 3D GT label.

1 Texture Generation

We utilized the TEXTure [3] to create 80 realistic UV texture maps for our DessiePIPE. We construct specific prompts for eight unique horse species by employing the format: *A photo of a* <SPECIES NAME>*, {} view.*

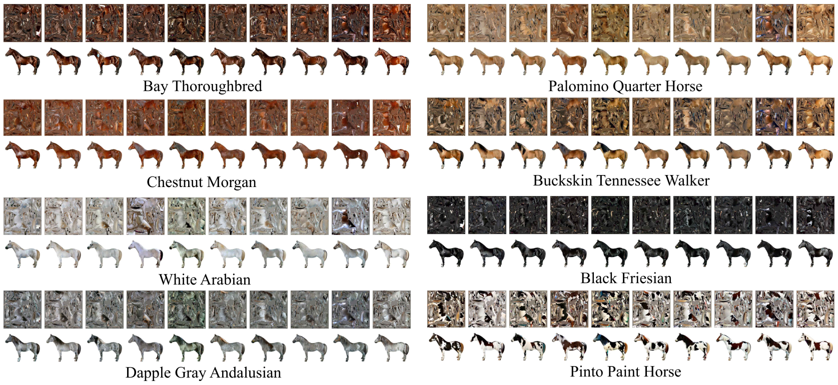

Fig. 1: UV texture maps created using TEXTure with eight horse species and rendered with the hSMAL model in zero pose.

<SPECIES NAME> is substituted with specific names, including "Bay Thoroughbred", "Palomino Quarter Horse", "Chestnut Morgan", "Buckskin Tennessee Walker", "White Arabian", "Black Friesian", "Dapple Gray Andalusian", and "Pinto Paint Horse". For each prompt, {}*view* is filled in with "front", "left",

"back", "right", "overhead" and "bottom", enabling the creation of textures for the horse model from various viewpoints through rotation. For each species, ten texture maps are generated using ten different random seeds. The resulting texture maps are shown in Fig. 1.

2 Qualitative Results

We show more reconstruction results of images from different domains in Fig. 2.

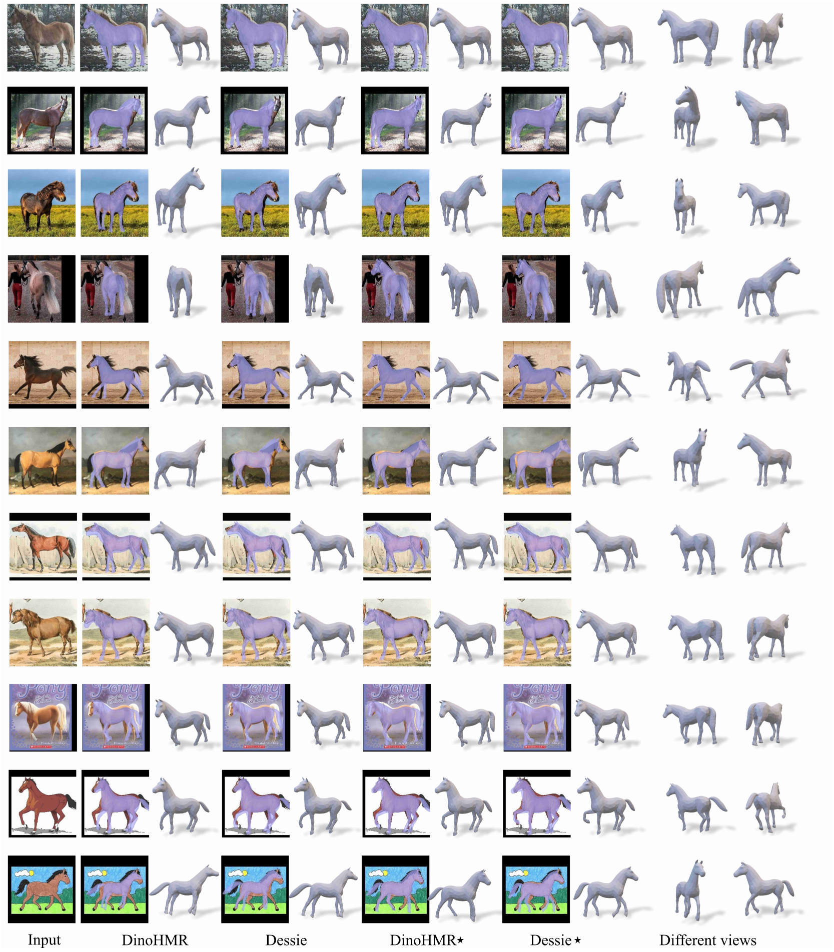

Fig. 2: More qualitative results of DinoHMR and Dessie before and after fine-tuning with real-world data. The first four rows showcase results from real-world images, while the subsequent rows are for images of horses from various domains, such as oil paintings and cartoons.

3 DINO Key Feature

The DINO features in the DinoHMR, visualised in the same way we mention in the main paper (Section 4.3), also focus on the subject.

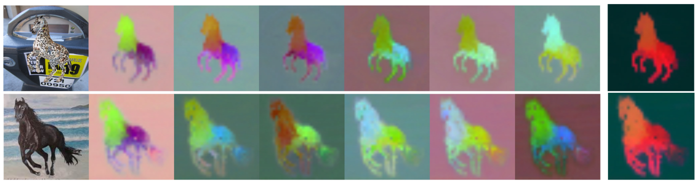

Fig. 3: Visualization of the leading PCA components of DinoHMR key features.

4 GT Label Effect

We investigate the effect of incorporating 3D GT labels in training Dessie with DessiePIPE. More specifically, we evaluate the skeletal joints in five real-horse sequences from PFERD [2], using 3D metric PAMPJPE and 2D metric PCK@0.1. For each skeletal joint, we manually define the correspondence on the mesh. For the 3D metric, we compare the predicted results with the 3D hSMAL GT, and for the 2D metric, we project the mesh vertices and compare them with the 2D GT. The results in Table 1 show that 3D GT loss L_{gt} improves 3D but degrades 2D performance. This is potentially caused by the predicted weak perspective camera, which does not match the actual camera, resulting in a mismatch between the projected 3D points and the detected 2D ones, as noted in TokenHMR [1].

Table 1: 3D evaluation on PFERD for Dessie.

L_{gt}	PAMPJPE ↓	PCK@0.1 ↑
✓	**102.59**	0.90
-	119.36	**0.92**

✓ or - indicates inclusion or exclusion of the loss component in training.

References

1. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
2. Li, C., Mellbin, Y., Krogager, J., Polikovsky, S., Holmberg, M., Ghorbani, N., Black, M.J., Kjellström, H., Zuffi, S., Hernlund, E.: The poses for equine research dataset (pferd). Scientific Data **11**(1), 497 (2024)
3. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: ACM SIGGRAPH (2023)